DE LA RECHERCHE À L'INDUSTRIE

HPSS IN EMBEDDED STORAGE SYSTEMS AT CEA

HPSS USER FORUM 2016

www.cea.fr

HUF2016 – From August 28th to September 2nd – New York City, USA

DE LA RECHERCHE À L'INDUSTRIE

TOPICS

- Overall architecture
- HPSS systems
- Embedded architecture
- HPSS movers configuration
- Benchmarks
- Plans

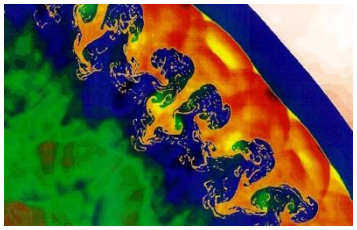
TERA, TGCC AND FRANCE GENOMIQUE

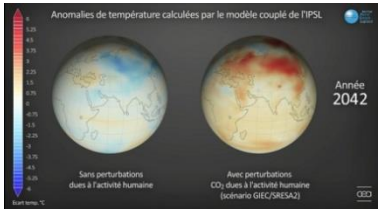
- 2 compute centers:
  - TERA for defense applications
  - TGCC for European research
    - hosting *France Génomique* (storage of DNA sequencing data)

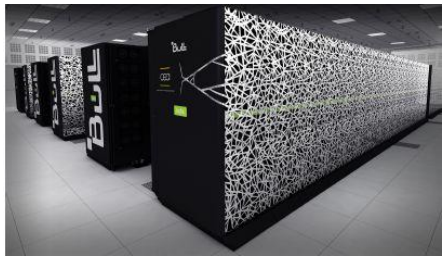
  ➔ **3 HPSS systems**

- Compute power:
  - TERA1K
    - P1: 2.586 Pflops
    - P2: 25 PFlops
  - TGCC/Curie: 3.2 PFlops
  - Fr. Génomique: ~100 Tflops

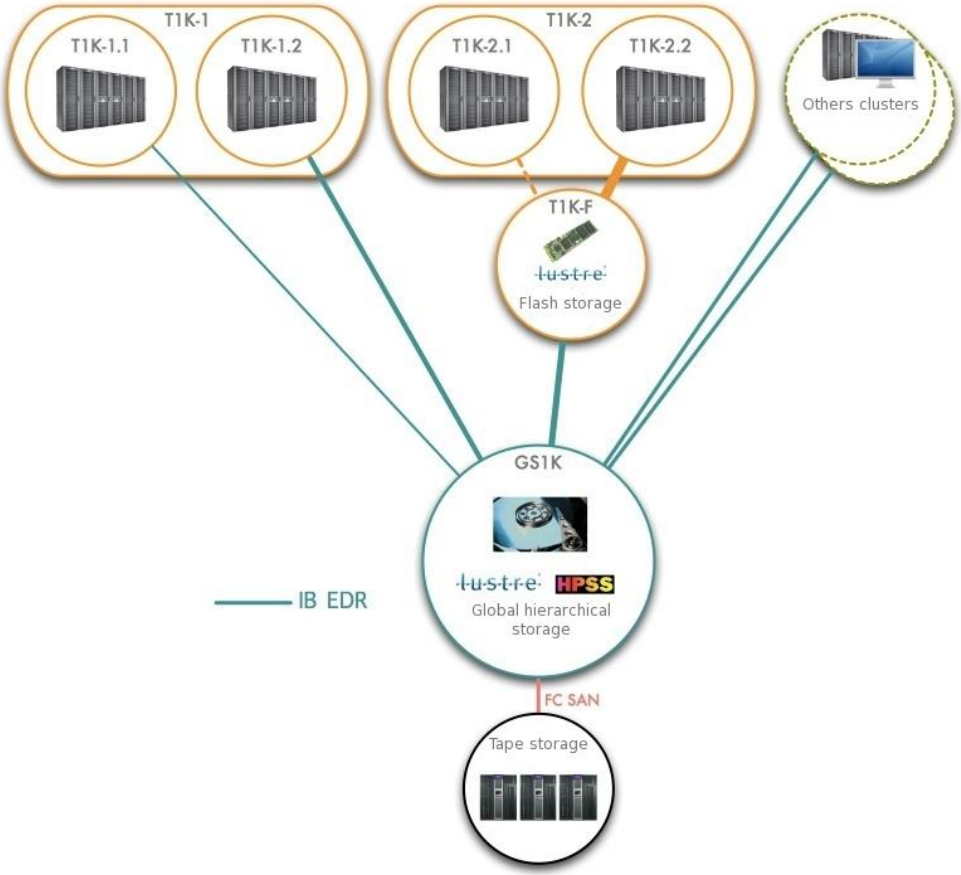
- 2 compute centers with a similar design:
  - Nearly the same architecture, technologies, tools and system software…







COMPUTE CENTERS ARCHITECTURE

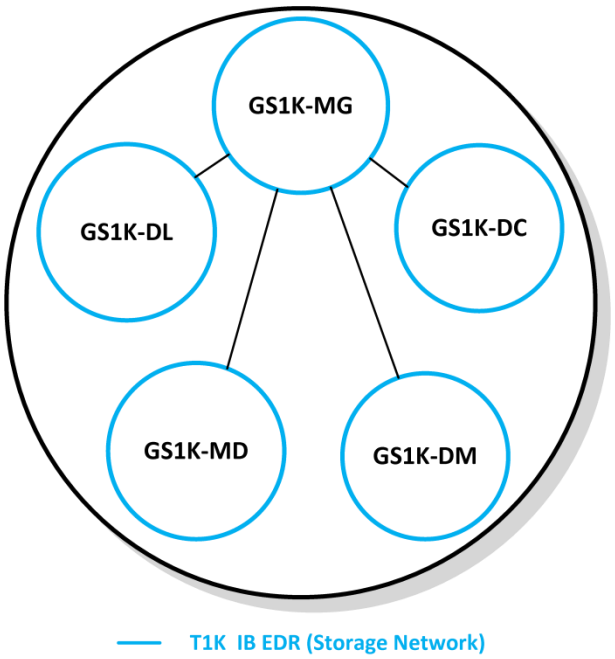


GS1K architecture

- User interface: Lustre/HSM (Lustre 2.5.3, upgrade to Lustre 2.7 in Q1 2017)
  - Seamless integration of HPSS as a Lustre backend

GLOBAL STORAGE ARCHITECTURE

Main idea



5 subsystems

Subsystems	Roles	Detail
GS1K-DL	Scratch Lustre FS (Fastest)	~530 GB/s, 10 PB
GS1K-DC	Store Lustre / HPSS	~225 GB/s, 30 PB
GS1K-MG	Management servers for the storage cluster	Spoms, Subnet managers, NFS servers
GS1K-MD	Metadata servers : Lustre (MDS), HPSS (DB2), Robinhood (MariaDB)	High frequency cpu servers, NetApp E-series 5600 (72 SSDs 400 GB)
GS1K-DM	Data movers : Lustre Agents (HPSS copytool), Lustre routers, HPSS tape movers	27 servers + 7 tape movers

GLOBAL STORAGE ARCHITECTURE, WHICH GOALS ?

An agile storage cluster :

- Idea: put Lustre and HPSS servers in the same cluster
  - Same kind of hardware for HPSS core and Lustre MDS (MD part)
  - Same kind of hardware for HPSS movers and Lustre OSS (DC part)

- Expected benefits:
  - Single Point Of Management, monitoring, …
  - Homogeneous hardware
  - Allow reallocating disk resources between Lustre and HPSS depending on their respective needs.
  - Reduce the datacenter footprint

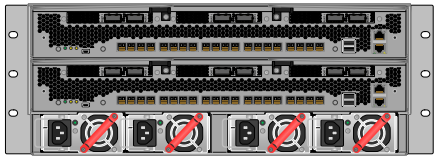
STORAGE SYSTEM DETAILS

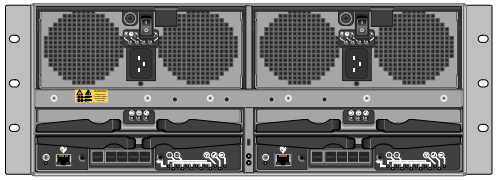
	Old system (2014 status)	New system
HPSS version	HPSS 7.3.3**p9b**	HPSS 7.4.3p2
OS	RHEL **6.4**	CentOS **6.7**
Disk level	3 NetApp E-5400 ~800 TB	1xDDN SFA14K-E 5 PB
Libraries	3 SL8500	3 SL8500
Tape techno	LTO5 (50 drives)	LTO5 (40 drives) T10KD (43 drives)
Lustre FS front-end	15 PB	20 PB
Stored in HPSS	34 PB	57 PB

EMBEDDED STORAGE FOR DISK LEVEL : SFA14K-E

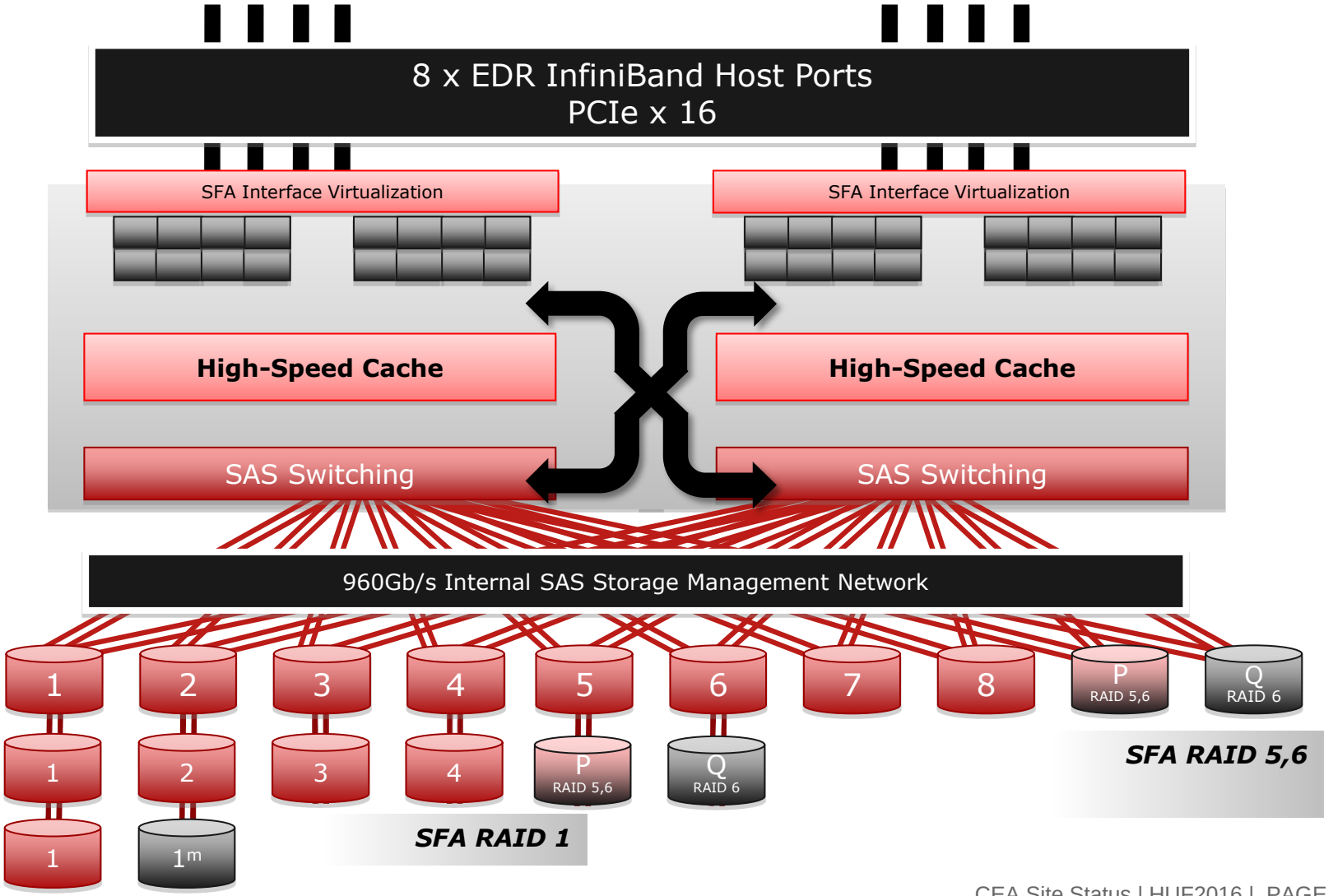
HPSS DISK LEVEL based on DDN SFA14K-E:

- x2 singlets with this configuration :
  - Bi-sockets Intel Haswell E5-2695
  - 128 GB RAM
  - 2 Mellanox Infiniband Connectx4 EDR dual ports
  - 2 Ethernet Gigabit ports
  - x4 Internal SSD Toshiba 480 GB
  - x5 SSD Toshiba 480 GB
- x810 Hitachi Hard Drives 8 TB (enclosures 8462)





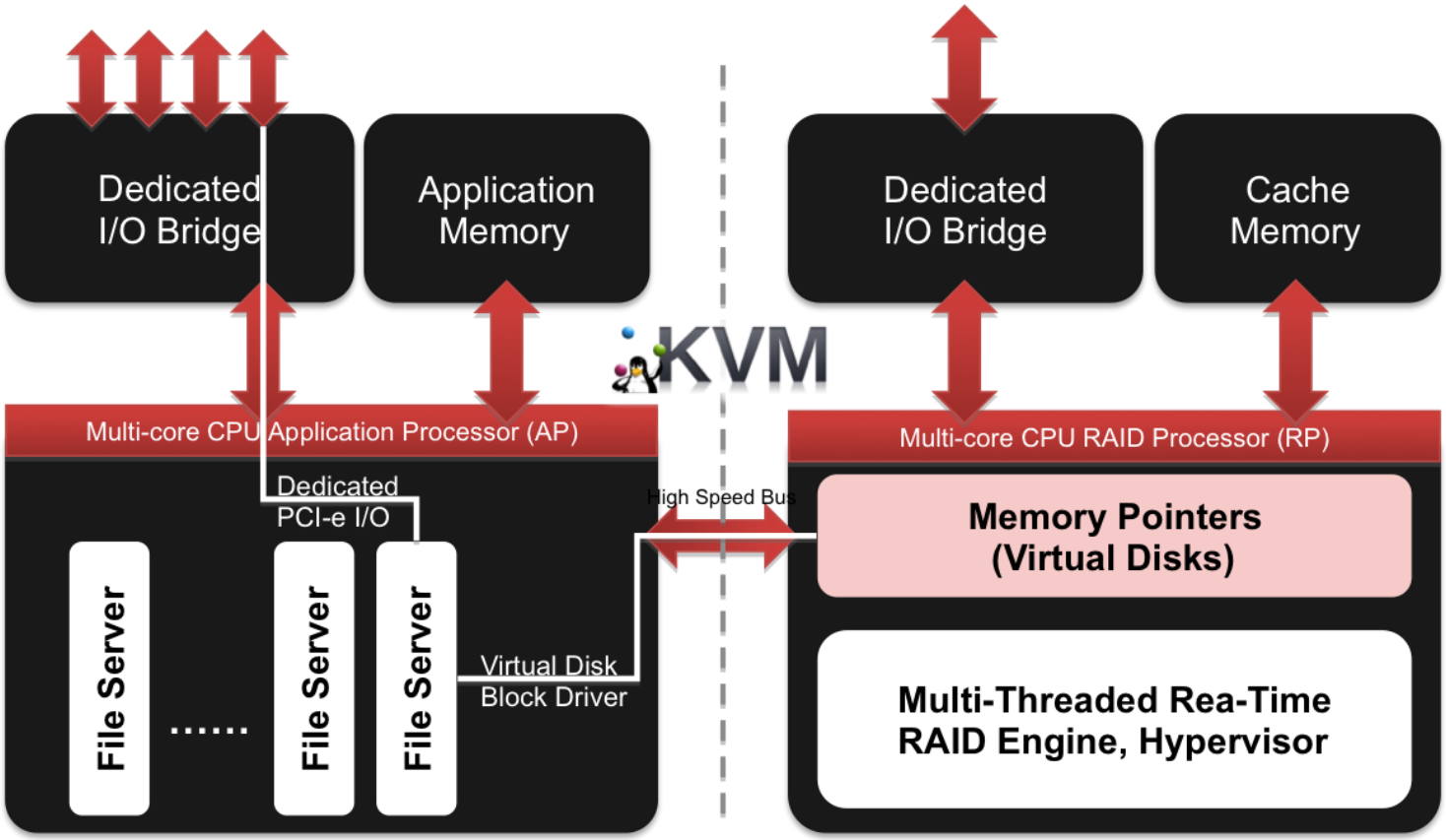
SFA14K-E : HOW IS INSIDE ?



SFA14K-E : WHAT CAN WE DO WITH ?

- **DDN SFA14K-E relies on KVM with a QEMU modified version**
- **Key features:**
  - RAID 1, 5 or 6 available for disks pool
  - up to 8 Virtual Machines
  - 48 cores to assign to VMs
  - 160 GB RAM to shared between VMs
  - PXE boot
  - SRIOV capabilities (Infiniband + Ethernet)
  - SFA Block driver (to « attach » enclosures disks to VMs)
- **DDN Terminology:**
  - POOL : disks raid group
  - VIRTUAL DISK (VD) : block device seen by VMs, part or a complete POOL
  - STACK : one VM
  - IOC : Virtual Function (Ethernet/Infiniband) = network interface seen by VM
  - IMAGE PATH : VM disk image

SFA14K-E : VM DIAGRAM

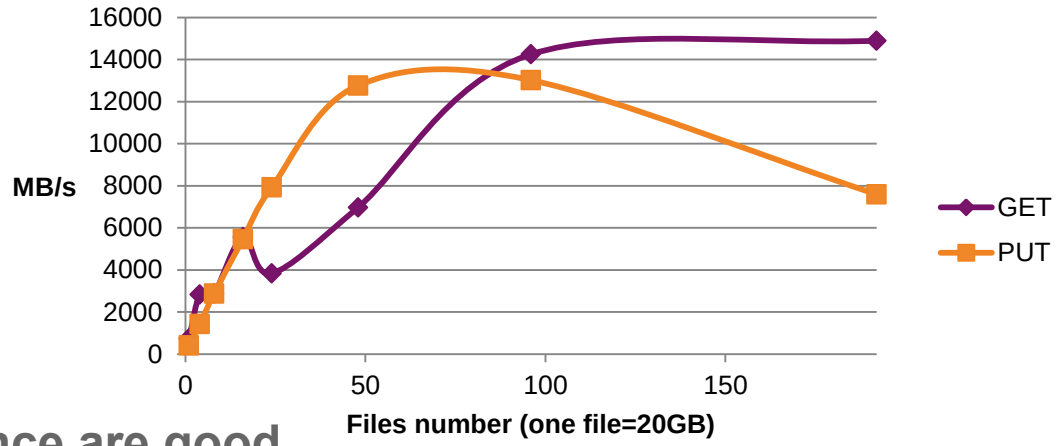


SFA14K-E : HPSS MOVERS CONFIGURATION

- **Key points for HPSS MOVERS :**
  - RAID 6 pool disks for enclosures 8462 => 80 pools of 8+2 (10 hotspares, one per enclosure)
  - 320 VD on the 80 pools => 3 BIG VD (19 712GB) and 1 SMALL VD (376GB)
  - 8 Virtual Machines => 8 HPSS disk movers (CentOS 6.7)
  - 6 Vcpu / mover
  - 20 GB RAM / mover
  - 1 dedicated Infiniband IOC per VM => 1 VM use 1 EDR link (100 Gbits/s)
  - 1 share Ethernet IOC by VM (mover administration)
  - 1 image path to store /boot
- **Details :**
  - All HPSS communications through Infiniband EDR (IPoIB)
  - /boot contains initram with SFA Block Driver module
  - Root (/) are on SSD drives
  - 40 devices / mover : 30 big, 10 small (distributed in 3 storage classes)

HPSS MOVERS EMBEDDED BENCHMARKS

- **Inputs:**
  - hpss_readlist/hpss_writelist
  - File size 20 GB
  - 10 clients with Infiniband EDR link (same Fabric as movers)



- **Performance are good …**
- Same hardware with Lustre, IOR benchmark => **33 GB/s** …(RDMA usage)
- But not bad, due to IPoIB overhead, CPU bound, pretty hard to go up to 70 Gbits/s per EDR link
- HPSS chooses devices automatically, sometimes, devices fall on the same pool => performance decrease

SUMMARY

- **SFA14-KE is a working platform:**
  - performance answers to our needs
  - The flexibility is here (we can grow Lustre FS or HPSS, easily)
  - Footprint is lower than before
  - homogeneous administration

- **Drawbacks**
  - Seems to be a SPOF but failover feature
  - Low footprint but density is not light (check your datacenter infrastructure and the floor strength)
  - Haswell limitation on heavyload, expect better performance with broadwell

PLANS FOR 2016-2017

- Q4 2016:
  - SFA14-KE tests with flash disks
  - SFA14-KE Passthrough feature (Latency impact ?)
  - RAIT
  - Upgrade HPSS movers to EL7 (7.4.3p3 migration)
- Q1 2017:
  - Upgrade to SFA14-KXE (Broadwell Processor)
  - HPSS 7.5.1 Validation
- Q2 2017:
  - Upgrade core server to EL7 (7.5.1 migration)
- Q3 2017:
  - Lustre store extension with Flash disks @**1TB/s**
- Q4 2017:
  - T1K Phase 2

END OF TAPE…

© Thomas Leibovici slide

Commissariat à l'énergie atomique et aux énergies alternatives
CEA / DAM Ile-de-France | Bruyères-le-Châtel - 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00

DAM Île-de-France

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019